

SECTOR IMPLEMENTATION GUIDE

GAISSF™ v1.0 Financial Services Sector Implementation Guide

Implementation guidance for the Global AI Safety and Security Framework

GAISSF-SEC-038 | Informative sector guidance | Draft for Publication
Version 1.0 | Public | External Review Open | Updated 6 October 2026

GAISSF™ Financial Services Sector Implementation Guide

GAISSF™ — Global AI Safety and Security Framework

Public Draft — External Review Open. Not Final Publication. Final Publication remains subject to qualified legal/regulatory, cybersecurity/model-risk, financial-crime/conduct/privacy/resilience, and explicit release-authorization review.

Executive summary

Financial-services AI can affect the balance sheet, customer access, financial crime exposure, market integrity, regulatory records, operational resilience, and confidence in critical services. GAISSF-SEC-038 translates GAISSF™ into operational implementation practices for financial institutions and material service providers.

This guide is implementation guidance, not a new conformance standard. Success means the organization can identify material AI use, assign accountability, constrain authority, validate outcomes, monitor degradation, preserve reconstructable evidence, respond to incidents, protect customers, and demonstrate resilience.

GAISSF baseline and applicability

GAISSF-SEC-038 does not add, remove, rename, or relax any GAISSF™ control. The authoritative baseline remains 59 controls across nine domains. Profile membership is defined in GAISSF-NOR-001. The Applicable-Control Baseline contains all 52 controls in D1-D8 plus the seven D9 controls where Physical AI or cyber-physical actuation is in scope. D1-D8 controls are not removed through sector tailoring. Operational and Optimized use the same applicable control scope; they increase operating-effectiveness evidence, monitoring, and assurance expectations rather than changing control membership.

Domain	Canonical GAISSF™ domain name	Scope note
D1	Model Integrity & Adversarial Robustness	Applies to all in-scope AI systems
D2	Runtime Security & Adversarial Defense	Applies to all in-scope AI systems
D3	Agentic Risk & Autonomous System Security	Applies to all in-scope AI systems
D4	Supply Chain & Third-Party AI Security	Applies to all in-scope AI systems
D5	Content Safety & Output Integrity	Applies to all in-scope AI systems
D6	Governance, Accountability & Human Oversight	Applies to all in-scope AI systems
D7	Human & Societal Harms	Applies to all in-scope AI systems
D8	Regulatory Alignment & Compliance	Applies to all in-scope AI systems
D9	Physical AI Safety	Applies wherever Physical AI or cyber-physical actuation is in scope

Where Physical AI or cyber-physical actuation is in scope, GAISSF™ D9 should be used alongside **PAI-SF™ v1.0** as applicable. **CSX-PAISF-004 is a mapping only; correspondence does not establish equivalence, evidence sufficiency, or automatic conformance.**

Executive implementation principles

Accountability remains with the financial entity

Outsourcing a model, platform, data source, or operational process does not transfer accountability for customer outcomes, financial crime controls, prudential risk, market integrity, resilience, or compliance.

Materiality classification reflects financial consequence

Classification must consider direct loss, restitution, conduct harm, financial crime exposure, regulatory action, capital and liquidity effects, market disruption, and correlated or systemic propagation.

Security and model risk operate as one assurance system

Threat modelling, model validation, secure engineering, data governance, operational resilience, change control, monitoring, and incident response must exchange evidence and escalation signals.

Human oversight must be operationally effective

A nominal reviewer is not effective oversight. The reviewer must have sufficient information, competence, time, authority, and a technically available intervention path before irreversible action.

Evidence must be decision-relevant and classified

Institutions should distinguish Claim Evidence Classes CEC-A Primary Verified, CEC-B Secondary Verified, CEC-C Simulation or Test, and CEC-D Anecdotal or Unverified. These local claim-evidence classes do not redefine ODA3's controlling T1-T4 evidence-sourcing taxonomy. Material assurance conclusions should not rely solely on CEC-D provider claims or policy statements.

Customer-impact controls are first-class

Controls must address discriminatory outcomes, unsuitable recommendations, wrongful denial, opaque adverse decisions, manipulation, exclusion, inaccessible challenge, and delayed remediation.

Proportionality does not remove minimum safeguards

Smaller institutions may combine governance forums and evidence processes, but critical actions, customer-impact decisions, financial crime controls, and critical-function dependencies still require explicit accountability and tested safeguards.

Translating AI risk to financial impact

Financial-services risk category	Financial impact vectors	Executive lens	Minimum control emphasis
FS-01 Unauthorized transactions	Direct monetary loss; customer restitution; fraud losses; liquidity leakage; operational investigation cost	CFO / CRO / CISO	Transaction limits, dual authorization, beneficiary controls, deterministic policy checks, immutable execution logs
FS-03 Financial crime degradation	Regulatory penalties; remediation programmes; correspondent banking restrictions; criminal exposure; customer exclusion	MLRO / CRO / Board Risk Committee	Independent validation, typology coverage, list freshness, alert integrity, override analytics, case reconstruction
FS-04 Credit and underwriting harm	Credit losses; conduct remediation; fair-lending exposure; capital-model distortion; reputational damage	CRO / Chief Credit Officer	Feature governance, calibration, fairness testing, adverse-action reasons, macro stress, exception monitoring
FS-05 Market conduct and integrity	Trading losses; market abuse findings; fines; licence restrictions; client claims; market disruption	Head of Markets / Compliance / CRO	Pre-trade limits, kill switches, surveillance, suitability checks, conflict controls, post-trade reconstruction
FS-06 Operational resilience failure	Service interruption; SLA penalties; lost revenue; customer compensation; supervisory intervention; contagion	COO / CIO / CISO	Critical-function mapping, fallback, capacity tests, RTO/RPO, dependency scenarios, tested recovery
FS-07 Third-party concentration	Correlated outage; exit cost; stranded data; supervisory access failure; vendor lock-in	CIO / Procurement / TPRM / CRO	Concentration analysis, audit rights, subcontractor visibility, portability, tested exit, alternate provider strategy
FS-12 Bias and exclusion	Customer remediation; litigation; enforcement; product withdrawal; loss of trust	Conduct Risk / Compliance / Product	Segment testing, outcome monitoring, accessible appeal, vulnerable-customer review, documented justification
FS-15 Agentic autonomy	Compounding transaction loss; unauthorized system change; cascading control bypass; prolonged execution loops	CISO / CTO / COO / CRO	Tool allow-lists, agent identity, cumulative limits, circuit breakers, approval boundaries, chain-of-action logs
FS-17 Systemic and contagion effects	Liquidity stress; correlated market activity; common-mode failure; sector-wide service disruption	Board Risk Committee / CRO / Treasury	Common dependency analysis, stress scenarios, exposure caps, cross-system monitoring, coordinated response

Financial-services AI risk categories

ID	Financial-services risk category	Interpretation	Inherent severity
FS-01	Unauthorized transaction or instruction generation	Fraudulent payments, orders, beneficiary changes, account actions, or privileged operator commands are generated, altered, or executed through AI compromise.	Critical
FS-02	Model manipulation and adversarial input	Evasion, poisoning, prompt injection, feature manipulation, synthetic identities, and feedback-loop exploitation alter risk decisions or tool behaviour.	Critical
FS-03	Financial crime control degradation	False negatives, alert suppression, typology drift, biased prioritisation, list staleness, or untraceable reasoning weakens AML, sanctions, fraud, or transaction monitoring.	Critical
FS-04	Credit, pricing and underwriting harm	Unfair, unstable, poorly calibrated, insufficiently explainable, or weakly governed lending, pricing, limit, affordability, collections, or insurance outcomes.	High
FS-05	Market conduct and integrity failure	AI contributes to manipulation, unsuitable advice, misleading communication, collusion, information leakage, uncontrolled autonomous execution, or surveillance failure.	Critical
FS-06	Operational resilience failure	AI dependency causes outage, processing error, capacity collapse, recovery failure, or inability to continue a critical or important function.	Critical
FS-07	Third-party and concentration risk	Shared model, cloud, identity, data, or platform dependencies create correlated failure, hidden subcontracting, exit constraints, or supervisory access gaps.	High
FS-08	Confidentiality and data leakage	Customer, authentication, transaction, trading, supervisory, portfolio, or proprietary data is exposed through training, inference, logging, retrieval, support, or model extraction.	Critical
FS-09	Privacy and data-subject rights failure	AI processing impairs access, rectification, erasure, objection, minimisation, purpose limitation, lawful basis, or privacy-impact assessment obligations.	High
FS-10	Identity, authentication and impersonation	Deepfakes, voice cloning, document synthesis, synthetic identity, social engineering, or AI-assisted account takeover defeats identity controls.	Critical
FS-11	Explainability, contestability and reasoning failure	The entity cannot reconstruct material factors, provide usable reasons, support challenge, or evidence	High / Critical by context

ID	Financial-services risk category	Interpretation	Inherent severity
FS-12	Bias, exclusion and disparate impact	meaningful review and remediation. Unjustified outcome differences arise across protected or vulnerable groups, geographies, income bands, channels, disability status, or customer segments.	High
FS-13	Data quality and provenance failure	Incorrect, stale, unlawfully sourced, unrepresentative, tampered, or lineage-deficient data affects decisions, reporting, and monitoring.	High
FS-14	Model drift and regime change	Performance or calibration degrades through economic shocks, behavioural change, fraud adaptation, market structure change, or feedback effects.	High
FS-15	Agentic and multi-agent execution failure	Chained agents exceed authority, lose identity continuity, compound tool-call limits, create orphaned tasks, or propagate erroneous plans across systems.	Critical
FS-16	Regulatory reporting and recordkeeping error	AI-generated returns, disclosures, books and records, model documentation, or supervisory responses are inaccurate, incomplete, or not reproducible.	High / Critical by context
FS-17	Systemic and contagion effects	Common models, correlated strategies, automated withdrawals, shared data dependencies, or synchronized risk responses amplify sector-wide stress.	Critical

Materiality assessment framework

Factor	Assessment question	Higher-impact indicator
Customer financial consequence	Can the system approve, deny, price, limit, collect, advise, freeze, or otherwise materially affect a customer?	Higher materiality where adverse action is difficult to reverse or affects essential financial access.
Authority and irreversibility	Can it move funds, execute trades, change entitlements, deploy code, or alter records?	Critical where action is autonomous, high-value, high-velocity, or irreversible.
Critical-function dependency	Does failure impair a critical or important business service?	Higher materiality where fallback is unavailable, slow, or untested.
Regulatory and evidentiary significance	Does it create regulatory reports, books and records, suspicious activity decisions, or supervisory submissions?	Higher materiality where reconstruction, reason-giving, or retention is mandatory.
Scale and concentration	How many customers, products, legal entities, markets, and third parties depend on it?	Higher materiality where one provider or model creates correlated exposure.
Adversarial attractiveness	Would manipulation produce fraud, evasion, access, market, or intelligence value?	Higher materiality where attackers can cheaply probe or influence inputs.
Vulnerability and fairness	Does it affect protected, vulnerable, thin-file, distressed, or digitally excluded groups?	Higher materiality where harm is difficult to detect through aggregate metrics.
Systemic propagation	Can outputs influence liquidity, common strategies, market pricing, or other institutions?	Critical where correlated response or feedback loops are plausible.

Representative use cases

ID	Use case	Indicative materiality	Accountable functions	Principal focus
UC-01	Fraud detection and transaction monitoring	Critical	Fraud Operations / CISO / Model Risk	False-negative risk, adversarial adaptation, alert integrity, override abuse
UC-02	AML and sanctions screening	Critical	MLRO / Financial Crime / Model Risk	Typology drift, list freshness, explainability, alert suppression, customer harm
UC-03	KYC, onboarding and identity verification	High	MLRO / Operations / CISO	Synthetic identity, deepfakes, document fraud, accessibility, manual escalation
UC-04	Credit scoring and underwriting	High	CRO / Chief Credit Officer / Compliance	Fairness, reason codes, calibration, affordability, macroeconomic regime change
UC-05	Insurance pricing and claims	High	Chief Underwriting Officer / Claims / Compliance	Discrimination, denial, fraud trade-offs, vulnerable customers, contestability
UC-06	Investment advice and portfolio construction	High	CIO / Conduct Risk / Compliance	Suitability, hallucinated facts, conflicts, product governance, volatility
UC-07	Algorithmic and AI-assisted trading	Critical	Head of Trading / Market Risk / Compliance	Runaway execution, manipulation, correlated strategies, kill-switch latency
UC-08	Customer service and financial guidance	High	Operations / Conduct Risk / CISO	Misleading advice, authentication, prompt injection, escalation failure
UC-09	Customer segmentation and marketing	High	Marketing / Conduct Risk / Privacy	Manipulation, vulnerable customers, prohibited proxies, consent, exclusion
UC-10	Collections and recoveries	High	Collections / Compliance / Customer Outcomes	Vulnerability, coercion, unfair prioritisation, prohibited communications
UC-11	Treasury, liquidity and capital analytics	Critical	CFO / Treasury / Prudential Risk	Forecast error, stress assumptions, concentration, reporting integrity
UC-12	Regulatory reporting and compliance drafting	High	Compliance / Finance / Legal	Accuracy, source traceability, approvals, retention, privileged information
UC-13	AI-assisted model validation	High	Model Risk / Internal Audit	Validator independence, automation bias, benchmark circularity, reproducibility
UC-14	Software engineering copilots	High	CTO / CISO	Insecure code, secret leakage, dependencies, unreviewed production change
UC-15	Cybersecurity operations	High	CISO / SOC	Automated containment error, attacker prompt injection, evidence integrity, excessive permissions

Claim evidence classes and quality

Claim Evidence Class	Definition	Examples	Limitation
CEC-A - Primary Verified	Direct, contemporaneous evidence from the control or system	Immutable logs, enforced policy decisions, signed approvals, versioned configurations, transaction blocks, access-control records	Strongest evidence of operation; still requires scope and integrity validation.
CEC-B - Secondary Verified	Independent or controlled review of primary evidence	Independent validation reports, internal audit testing, reconciliations, quality assurance sampling, third-party assurance with verified scope	Useful for design and operating effectiveness when independence and sampling are clear.
CEC-C - Simulation / Test	Evidence generated under designed test conditions	Red-team results, resilience exercises, tabletop simulations, synthetic fairness tests, failover tests	Supports plausibility and preparedness; does not prove sustained production effectiveness.
CEC-D - Anecdotal / Unverified	Assertions without sufficient independent corroboration	Provider marketing claims, policy-only statements, interviews without records, screenshots without provenance	May identify leads; should not carry material assurance conclusions alone.

Quality	Meaning
None	No evidence supplied or evidence does not address the control objective.
Weak	Policy or assertion exists, but scope, provenance, approval, timing, or operating evidence is missing.
Partial	Evidence covers design or selected operation, but material populations, periods, exceptions, or dependencies are incomplete.
Strong	Evidence is current, attributable, complete for the stated scope, independently challenged where required, reproducible, and linked to exceptions and remediation.

Lifecycle implementation guidance

1. Initiation and approval

- Define purpose, prohibited outcomes, affected customers, legal entities, critical functions, and intended authority.
- Assign accountable executive, business owner, model owner, data owner, security owner, control owners, independent validators, and incident escalation authority.
- Ensure accountable executives and control owners have role-appropriate AI literacy.
- Classify inherent risk before production data access, procurement commitment, or pilot authority.

2. Data acquisition and preparation

- Establish provenance, legal basis, licensing, purpose, representativeness, quality, lineage, retention, and access controls.
- Test for sensitive-data leakage, prohibited attributes and proxies, temporal contamination, label error, manipulation, and economic-regime bias.
- Design and test privacy rights, minimisation, deletion, and correction processes where applicable.

3. Design and development

- Apply secure architecture, threat modelling, segregation, secrets management, dependency governance, and abuse-case testing.
- Constrain tools, identities, actions, transaction values, destinations, code execution, network access, retrieval, and delegation.
- Design reconstructable records and effective human intervention points before implementation.

4. Independent validation and challenge

- Independent validation requires sufficient organisational, reporting, intellectual, and access independence to reach and escalate an adverse conclusion.
- Validate conceptual soundness, data suitability, implementation correctness, calibration, discrimination, robustness, explainability, stress performance, and limitations.
- Record unresolved findings, conditions, compensating controls, expiry dates, and residual-risk acceptance.

5. Deployment and change control

- Use controlled release, versioning, approvals, phased rollout where appropriate, rollback, and predefined stop conditions.
- Govern changes to model, prompt, retrieval corpus, agent, tool, provider, feature, threshold, policy, data, hosting, and identity.
- Prevent unmanaged employee use and shadow AI in sensitive workflows.

6. Production monitoring

- Monitor technical, customer, fairness, fraud, financial-crime, security, operational, data, economic-regime, and concentration indicators.
- Tie every material threshold to ownership, action, severity, escalation, customer protection, and notification analysis.
- Review complaints, losses, appeals, overrides, near misses, and incidents as model-risk signals.

7. Incident response and remediation

- Use the GAISSF-SEC-038 incident taxonomy and integrate cyber, fraud, privacy, conduct, financial-crime, resilience, reporting, and crisis processes.
- Preserve model, data, prompt, policy, agent, tool, approval, identity, and transaction evidence.
- Assess customer remediation, reversal, suspension, disclosure, notification, and systemic propagation.

8. Retirement and exit

- Define archival, retention, access revocation, data return/deletion, customer-impact review, dependency removal, and provider exit.
- Test portability and continuity before a critical dependency becomes irreversible.
- Continue monitoring delayed complaints, claims, corrections, and reporting consequences.

9. Post-implementation review

- At 6-12 months or earlier when triggered, review control effectiveness, assumptions, realised outcomes, incidents, exceptions, concentration, and unintended effects.
- Document lessons learned, required control changes, risk re-approval, and improvement ownership.

Illustrative monitoring metrics

Category	Illustrative metrics	Limitation
Technical discrimination and ranking	AUC-ROC where suitable; precision-recall; F1; false-positive/false-negative rates; calibration error; top-k recall	Do not rely on a single aggregate metric; choose metrics consistent with class imbalance and decision cost.
Customer outcomes	Approval/denial rates; pricing dispersion; complaint rate; appeal success; remediation value; time to resolution	Segment by relevant products, channels, geographies, vulnerability indicators, and protected groups where lawful.
Fairness and exclusion	Disparate impact ratio; equal opportunity difference; calibration by group; error-rate parity; accessibility failure rate	Metric selection requires legal and contextual review; no single fairness metric establishes fairness.
Fraud and financial crime	Detection rate; confirmed miss rate; alert-to-case conversion; typology coverage; list freshness; override concentration	Use post-event investigations and emerging typologies, not only historical labelled data.
Operational resilience	p95/p99 latency; throughput; timeout rate; dependency failure rate; fallback activation; recovery time; queue growth	Tie thresholds to critical-service impact and documented actions.
Human oversight	Review completion; override rate; override reversal; escalation time; reviewer disagreement; intervention success	High review volume can indicate poor model quality; low override volume can indicate automation bias.
Agentic execution	Tool-call count; cumulative transaction value; denied actions; loop duration; orphaned tasks; identity handoff failures	Monitor the complete chain, not only individual agent steps.
Data and drift	Population stability; feature drift; missingness; lineage exceptions; concept drift; out-of-distribution rate	Pair statistical triggers with business and economic regime indicators.
Security	Prompt-injection detection; unauthorized tool requests; model extraction signals; secret exposure; anomalous access	Security events must feed model-risk and incident processes.

AI incident classification

Class	Category	Illustrative events
AI-1	Security compromise	Prompt injection, poisoning, model theft, unauthorized access, tool misuse, credential exposure
AI-2	Decision integrity failure	Incorrect, unstable, biased, unexplainable, or unreconstructable decision
AI-3	Financial crime control failure	Missed fraud, sanctions, AML, KYC, or alert integrity failure
AI-4	Customer or conduct harm	Mis-selling, unfair denial, manipulation, inaccessible service, or ineffective redress
AI-5	Operational resilience failure	Outage, capacity collapse, dependency failure, recovery or fallback failure
AI-6	Data and privacy failure	Leakage, unlawful processing, rights failure, provenance or minimisation breach
AI-7	Agentic execution failure	Excess authority, cascading actions, orphan loop, cumulative-limit bypass, unsafe tool chain
AI-8	Reporting and recordkeeping failure	Incorrect return, disclosure, books and records, missing lineage, or evidence loss
AI-9	Market or systemic event	Manipulative activity, correlated strategies, liquidity pressure, common-mode or contagion event

Sector interpretations and illustrative evidence

Area	Interpretation	Illustrative evidence
Governance and accountability	Board and senior management oversight should identify material AI-enabled financial services, accountable executives, risk appetite, prohibited uses, and independent challenge.	Board-approved risk appetite; accountable executive register; committee minutes; exception approvals; management information.
Inventory and classification	Maintain an inventory covering in-house, embedded, vendor, employee-adopted, and shadow AI; link it to critical functions and materiality.	AI register; architecture and data-flow maps; provider register; critical-function links; materiality rationale.
Threat modelling	Cover adversarial inputs, fraud adaptation, injection, extraction, poisoning, identity deception, agent misuse, and common-mode failure.	Threat model; abuse cases; attack trees; red-team scope; agent interaction tests; residual-risk record.
Data governance and privacy	Control provenance, lawful use, rights, representativeness, quality, lineage, access, minimisation, retention, and contamination.	DPIA; data sheets; lineage; quality reports; rights workflow tests; feature approvals; retention evidence.
Independent validation	Validate conceptual soundness, implementation, fairness, calibration, stability, robustness, explainability, and stress performance with defined independence.	Validation charter; independent report; benchmark and stress results; challenge log; limitations; approval conditions.
Secure engineering and AI lifecycle management	Integrate AI components, prompts, retrieval, tools, agents, and data pipelines into secure SDLC and controlled lifecycle gates.	Architecture review; AIBOM/SBOM; code review; pipeline controls; version registry; release/retirement approvals.
Procurement and vendor management	Assess provenance, subcontractors, data use, security, validation, change, resilience, audit rights, supervisory cooperation, concentration, and exit.	Due diligence; contract clauses; assurance scope review; subcontractor register; concentration assessment; exit test.
Access, authority and multi-agent control	Apply least privilege, workload identity, delegation provenance, cumulative limits, circuit breakers, and segregation of duties.	RBAC/ABAC; signed delegation; tool allow-list; transaction limits; circuit-breaker tests; chain logs.
Human oversight	Provide effective supervision and intervention; define human review points and monitor automation bias and override patterns.	Interface screenshots; reviewer competence; authority matrix; override logs; escalation tests; QA sampling.
Monitoring and drift	Monitor technical performance, customer outcomes, fairness, fraud	Dashboards; metric definitions; thresholds; alerts; trend analysis;

Area	Interpretation	Illustrative evidence
	adaptation, calibration, latency, data/concept drift, and concentration.	recalibration and stop decisions.
Incident response	Integrate AI incidents with cyber, fraud, conduct, privacy, financial-crime, reporting, resilience, and customer remediation.	Taxonomy; playbooks; decision tree; evidence preservation; notifications; customer remediation; lessons learned.
Resilience and recovery	Provide tested fallback, degraded mode, capacity, recovery objectives, manual alternatives, and dependency-aware scenarios.	BIA; RTO/RPO; failover and capacity test; fallback exercise; dependency map; recovery evidence.
Customer transparency and redress	Support accurate disclosures, material reasons, challenge, correction, escalation, accessibility, and timely remediation.	Customer notices; reason validation; appeal records; accessibility tests; remediation metrics.
Recordkeeping and reproducibility	Reconstruct material decisions, versions, data/features, prompts, policy states, agents, tools, approvals, and operator actions.	Version records; immutable logs; decision trace; action-chain history; retention controls; audit extracts.

Specific implementation patterns

Fraud, AML, sanctions and KYC

- Separate model detection performance from alert handling, case management, list management, and investigator behaviour; each can fail independently.
- Estimate false negatives using confirmed events, look-backs, post-event investigations, emerging typologies, adversarial simulations, and external intelligence, not only labelled historical data.
- Protect features, lists, rules, thresholds, investigator notes, feedback channels, and case dispositions from unauthorized change and poisoning.
- Maintain reason codes and case reconstruction sufficient for internal challenge and applicable reporting; do not assume one explanation technique satisfies every legal context.
- Monitor de-risking, freezing, closure, escalation, and override outcomes for disproportionate customer harm and control evasion.

Credit, pricing, underwriting and collections

- Define permissible features and proxies, reason codes, fairness metrics, affordability logic, approval thresholds, exception routes, and adverse-decision processes before release.
- Validate across protected and vulnerable groups, thin-file populations, products, channels, geographies, and economic scenarios.
- Monitor calibration and realised outcomes through rate, employment, inflation, property, catastrophe, and liquidity changes.
- Test whether human overrides reduce harm or introduce inconsistent, undocumented, or discriminatory outcomes.

Trading, investment and market activity

- Set instrument, venue, order type, size, price, loss, position, exposure, latency, and market-condition limits outside the AI model.
- Implement deterministic kill switches and manual control that remain available during model, data, network, venue, or identity degradation.
- Assess manipulation, correlated strategy, information leakage, suitability, conflict, and hallucinated-market-fact risk.
- Reconstruct strategy/model version, market data, prompts, orders, cancellations, overrides, surveillance alerts, and approvals.

Generative AI and customer interactions

- Use approved grounding sources, explicit uncertainty handling, prohibited-content rules, escalation paths, and authenticated account-action boundaries.
- Protect retrieval stores, prompts, memory, tools, and conversation logs from injection, poisoning, disclosure, and cross-customer leakage.
- Test vulnerable-customer, accessibility, language, bereavement, scam, fraud, distress, complaint, and dispute scenarios.
- Disclosure that AI is used may support transparency but does not substitute for accuracy, safety, suitability, and redress.

Agentic and multi-agent workflows

- Assign each agent a unique workload identity and verify agent-to-agent authentication, delegation, provenance, and revocation across the complete chain.
- Apply both per-step and cumulative limits for transaction value, tool calls, execution time, data access, destinations, code changes, and retries; downstream agents must not reset upstream limits.

- Separate planning, approval, and execution agents where feasible; require independent authorization for irreversible, privileged, or high-value actions.
- Prevent orphaned execution by using leases, heartbeats, bounded retries, idempotency controls, compensating actions, and deterministic completion states.
- Use a circuit breaker independent of the agent stack that can halt anomalous behaviour using deterministic rules and can be tested without model cooperation.
- Log the entire plan and action chain: agent identity, delegation, prompt/context, policy decision, tool parameters, credential context, result, exception, approval, and rollback.
- Test cascading misinformation, conflicting agent goals, recursive delegation, shared-memory poisoning, partial completion, and control failure under degraded dependencies.

AI-assisted model validation

- Prevent circular validation in which the validator uses the same model family, training data, benchmark, or provider assumptions without disclosure and compensating challenge.
- Require human validators to understand generated analyses, reproduce material tests, and own conclusions rather than accepting automated reports.
- Record validator tool versions, prompts, source data, code, assumptions, exceptions, and independent challenge.

Software engineering and cyber operations

- Constrain copilots and agents from committing, deploying, rotating credentials, changing firewall or identity policy, or containing assets without approved authority.
- Require code review, testing, dependency controls, secret scanning, and environment segregation for AI-generated changes.
- Preserve incident evidence and prevent attacker-controlled content from directly invoking response tools.

Illustrative implementation roadmap (non-prescriptive)

The timeframes below are planning aids only. They must be adapted to system authority, customer and prudential consequence, critical-service dependencies, regulatory obligations, operating constraints, and the organisation's existing control environment.

Timeframe	Objective	Minimum activities	Exit evidence
0-30 days	Establish accountability and exposure baseline	Approve interim governance; inventory material AI; identify critical functions and high-risk actions; constrain unmanaged deployments; define incident route.	Executive sponsor; inventory coverage; critical gaps; prohibited uses; peer-review status recorded.
31-60 days	Prioritise controls and evidence	Establish the Applicable-Control Baseline; use the technically reconciled GAISSE cross-reference as implementation guidance; complete threat and data reviews; assess critical providers.	Materiality-prioritized remediation plan; evidence owners; mapping review actions; concentration and contractual issues.
61-90 days	Test high-consequence systems	Perform independent validation, adversarial tests, resilience scenarios, circuit-breaker and fallback exercises, customer-outcome review, and incident simulation.	Critical findings resolved or accepted; stop mechanisms tested; residual risk approved.
3-6 months	Integrate the operating model	Embed lifecycle gates, procurement clauses, change governance, monitoring, complaint feedback, assurance sampling, and role-based AI literacy.	Repeatable controls; management information; traceable approvals; policy/process integration.
6-12 months	Review effectiveness and mature	Conduct post-implementation review, concentration and contagion analysis, thematic reviews, control-effectiveness testing, and independent assessment-readiness review.	Sustained effectiveness; lessons learned; evidence quality improvement; assurance report with limitations.

Common failure patterns

- Treating AI governance as a policy exercise without technical enforcement, deployment inventory, and operating evidence.

- Validating predictive accuracy while omitting adversarial, resilience, fairness, customer-outcome, fraud-adaptation, and extreme-event testing.
- Using provider documentation as a substitute for entity-specific threat assessment, validation, and exit testing.
- Placing human review after the transaction, communication, or customer impact has become irreversible.
- Failing to govern prompts, retrieval corpora, tools, agent identities, thresholds, features, and provider updates as controlled changes.
- Monitoring averages that conceal severe subgroup, product, channel, region, or tail-event failures.
- Keeping kill switches, circuit breakers, or fallback procedures that are untested, inaccessible, or too slow.
- Recording only the final output and losing the model version, input context, policy state, delegation chain, tool actions, and approvals.
- Treating complaints, fraud losses, near misses, overrides, and operational incidents as separate from model monitoring.
- Ignoring concentration and correlated failure across shared models, cloud, identity, market data, payment rails, and data providers.
- Assuming a low-volume system is low risk when it can move funds, alter rights, create regulatory records, or affect a critical function.

Notably Absent

- No claim that GAISSF-SEC-038 establishes legal or regulatory compliance.
- No universal list of applicable regulators, laws, supervisory expectations, or reporting thresholds.
- No presumption that explainability alone makes a model safe, fair, lawful, or suitable.
- No presumption that provider assurance replaces the financial entity's own assessment, validation, and monitoring.
- No blanket prohibition or endorsement of generative, predictive, or agentic AI.
- No fixed materiality or residual-risk threshold applicable to every financial entity.
- No claim that GAISSF-SEC-038 mapping establishes equivalence, evidence sufficiency, or automatic GAISSF conformance; independent specialist review remains open.
- No assertion that human review is effective without evidence of timing, competence, information, authority, and intervention capability.
- No assumption that successful model validation eliminates cyber, fraud, conduct, privacy, resilience, or systemic risk.
- No representation that this guide supersedes law, regulation, supervisory direction, contract, or the GAISSF normative standard.
- No claim that the source register is exhaustive or that every listed source applies to every entity or jurisdiction.

GAISSF cross-reference register

The cross-reference was technically reconciled by ODA3 against the locked GAISSF-NOR-001 / GAISSF-NOR-004 v1.0 baseline. It is implementation guidance, not a substitute for the Applicable-Control Baseline or an independent conformance determination. Independent specialist review remains open.

Ref	Implementation area	GAISSF v1.0 reference	Relationship	Status
XR-01	Governance and accountability	D6-CTL-01 to D6-CTL-07; D8-CTL-01 to D8-CTL-05	Direct and supporting	Technically reconciled; independent review open
XR-02	Inventory and classification	D4-CTL-01; D4-CTL-06; D6-CTL-03	Direct and supporting	Technically reconciled; independent review open
XR-03	Threat modelling	D1-CTL-01 to D1-CTL-09; D2-CTL-01 to D2-CTL-06; D3-CTL-01 to D3-CTL-07; D7-CTL-H05	Supporting	Technically reconciled; independent review open
XR-04	Data governance and privacy	D1-CTL-01; D3-CTL-06; D3-CTL-07; D5-CTL-02; D5-CTL-05; D5-CTL-06	Direct and supporting	Technically reconciled; independent review open
XR-05	Independent validation	D1-CTL-03; D2-CTL-03; D6-CTL-03	Supporting	Technically reconciled; independent review open
XR-06	Secure engineering and AI lifecycle management	D4-CTL-01; D4-CTL-02; D4-CTL-03; D4-CTL-07; D6-CTL-05; D8-CTL-05	Direct and supporting	Technically reconciled; independent review open
XR-07	Procurement and vendor	D4-CTL-03; D4-CTL-05; D6-CTL-06;	Direct and supporting	Technically reconciled;

Ref	Implementation area	GAISSE v1.0 reference	Relationship	Status
	management	D6-CTL-07		independent review open
XR-08	Access, authority and multi-agent control	D2-CTL-05; D3-CTL-01; D3-CTL-02; D3-CTL-03; D3-CTL-05	Direct	Technically reconciled; independent review open
XR-09	Human oversight	D6-CTL-01; D6-CTL-02	Direct	Technically reconciled; independent review open
XR-10	Monitoring and drift	D1-CTL-03; D3-CTL-03; D4-CTL-04; D6-CTL-07	Direct and supporting	Technically reconciled; independent review open
XR-11	Incident response	D6-CTL-04; D7-CTL-H04; D8-CTL-04	Direct	Technically reconciled; independent review open
XR-12	Resilience and recovery	D6-CTL-07; D9-CTL-02; D9-CTL-03	Direct and conditional	Technically reconciled; independent review open
XR-13	Customer transparency and redress	D5-CTL-01; D5-CTL-02; D5-CTL-05; D6-CTL-01; D6-CTL-02; D6-CTL-03	Supporting	Technically reconciled; independent review open
XR-14	Recordkeeping and reproducibility	D6-CTL-02; D6-CTL-03; D6-CTL-04; D6-CTL-05	Direct and supporting	Technically reconciled; independent review open

External source register

ID	Source	Issuer	Edition/status	Applicability	Use	Official location
SRC-01	Regulation (EU) 2022/2554 - Digital Operational Resilience Act (DORA)	European Union	Adopted 14 Dec 2022; applies from 17 Jan 2025	EU financial entities and ICT third-party risk; verify entity scope	Operational resilience, ICT risk, incidents, testing, third-party risk	Official EU text
SRC-02	Regulation (EU) 2024/1689 - Artificial Intelligence Act	European Union	Current consolidated version 27 Jul 2026; phased application	EU and extraterritorial applicability requires legal analysis	AI risk, high-risk systems, governance, literacy, transparency	EUR-Lex current consolidated text
SRC-03	Regulation (EU) 2016/679 - General Data Protection Regulation	European Union	Applicable since 25 May 2018	Personal data processing in EU/EEA and applicable extraterritorial cases	Privacy, lawful basis, rights, minimisation, automated decisions	Official EU text
SRC-04	Directive 2014/65/EU - MiFID II	European Union	Applicable framework; verify consolidated text and national implementation	Investment services and markets	Suitability, conduct, governance, records, algorithmic trading context	Official EU text
SRC-05	Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1	NIST	AI RMF 1.0 (Jan 2023); NIST states revision is underway in 2026	Voluntary cross-sector framework	AI lifecycle risk-management concepts	NIST AI RMF 1.0
SRC-06	Generative AI Profile,	NIST	July 2024	Voluntary profile	Generative-AI risk and	NIST GenAI

ID	Source	Issuer	Edition/ status	Applicability	Use	Official location
	NIST AI 600-1				action context	Profile
SRC-07	Principles for operational resilience	Basel Committee on Banking Supervision	31 March 2021	Banks; implementation depends on jurisdiction and supervisor	Operational resilience and critical operations	BCBS operational resilience
SRC-08	Principles for the sound management of operational risk	Basel Committee on Banking Supervision	Revised March 2021	Banks; implementation depends on jurisdiction and supervisor	Operational risk governance and control environment	BCBS operational risk
SRC-09	PCI Data Security Standard	PCI Security Standards Council	PCI DSS v4.0.1; current PCI DSS v4.x baseline as checked 6 Oct 2026	Entities within PCI DSS scope	Payment-card data security	PCI DSS
SRC-10	ISO/IEC 42001:2023 - Artificial intelligence - Management system	ISO/IEC	Edition 1, published Dec 2023; current as checked 6 Oct 2026	Voluntary management-system standard; external certification may be available independently of GAISST	AI management-system context	ISO/IEC 42001
SRC-11	ISO/IEC 23894:2023 - Artificial intelligence - Guidance on risk management	ISO/IEC	Edition 1, published Feb 2023; current as checked 6 Oct 2026	Voluntary guidance	AI risk-management context	ISO/IEC 23894
SRC-12	OECD AI Incidents Monitor	OECD	Living resource; verify current taxonomy and usage terms	Illustrative incident-learning resource	Incident examples and taxonomy context	OECD AI Incidents Monitor

Publication gate status

Gate	Status	Evidence / limitation
Artifact consistency	Closed / Verified after 6 Oct 2026 public-draft rebuild	DOCX, PDF, Markdown, JSON, workbook, cross-reference CSV, website summary/landing page, README and manifest/checksums reconciled for the controlled public-draft release.
External source bibliographic verification	Closed / Verified 6 Oct 2026	Official publisher sources rechecked for the public-draft edition; legal applicability remains outside this gate.
GAISST cross-reference technical reconciliation	Completed / Independent review open	Reconciled against locked GAISST-NOR-001 / GAISST-NOR-004 v1.0 control identities. Mapping is correspondence only and does not establish equivalence, evidence sufficiency, or automatic conformance.
Qualified legal and regulatory review	Open	Jurisdiction-specific applicability, legal conclusions, reporting duties, and article-level mappings require

Gate	Status	Evidence / limitation
Cybersecurity and model-risk peer review	Open	qualified review. Independent specialist review and named disposition remain required before Final Publication.
Financial-crime, conduct, privacy and resilience peer review	Open	Named specialist review and disposition of comments remain required.
Publication approval	Open - Final Publication only	Final Publication requires explicit authorized release approval after final-publication gates are closed or expressly accepted. Public draft publication does not represent Final Publication approval.

Publication history: originally prepared 1 May 2026; technically reconciled 5 October 2026; public-draft refinement issued 6 October 2026. Final publication date remains unset.

© 2026 ODA3 Pvt Ltd. Market-facing publisher: ODA3 Institute. Use governed by the GAISSE Ecosystem Licence (GEL) v1.0. GAISSE Ecosystem developed by ODA3 Institute, licensed under GEL v1.0. Enquiries: contact@oda3.org.